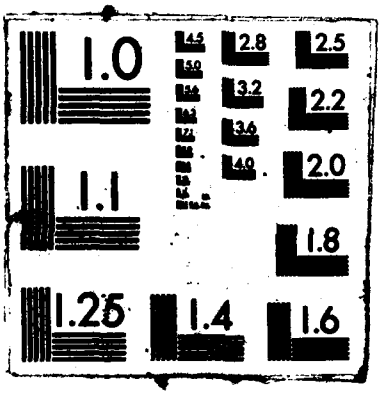
AD-A181 145 BINOMIAL N ESTIMATION A BAYES EMPIRICAL BAYES APPROACH 1/1
(U) WASHINGTON UNIV SEATTLE DEPT OF STATISTICS
A E RAFTERY JUL 86 TR-85 N00014-84-C-0169

UNCLASSIFIED F/G 12/3 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 85	2. GOVT ACCESSION NO. AD-A181 145	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Binomial N estimation: A Bayes empirical Bayes approach		5. TYPE OF REPORT & PERIOD COVERED TR 12/86 - 6/88
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Adrian E. Raftery		8. CONTRACT OR GRANT NUMBER(s) N00014-84-C-0169
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics, GN-22 University of Washington Seattle, WA 98195		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-661-003
11. CONTROLLING OFFICE NAME AND ADDRESS ONR Code N63374 1107 NE 45th Street Seattle, WA 98105		12. REPORT DATE July 1986
		13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Binomial N estimation; Bayes empirical Bayes

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A Bayes empirical Bayes approach to the problem of estimating N in the binomial distribution is presented. This provides a simple and flexible way of specifying prior information, and also allows a convenient representation of vague prior knowledge. In addition, it yields a solution to the interval estimation problem. The Bayes estimator corresponding to the relative squared error loss function and a vague prior distribution is shown to be stable, and to compare favorably with the estimators introduced/CONTINUED ...

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

by Olkin et al. (1981) and Carroll and Lombard (1985).

-A-

Binomial N Estimation: A Bayes Empirical Bayes Approach

Adrian E. Raftery

Department of Statistics,GN-22,
University of Washington,
Seattle, WA 98195.

ABSTRACT

A Bayes empirical Bayes approach to the problem of estimating N in the binomial distribution is presented. This provides a simple and flexible way of specifying prior information, and also allows a convenient representation of vague prior knowledge. In addition, it yields a solution to the interval estimation problem. The Bayes estimator corresponding to the relative squared error loss function and a vague prior distribution is shown to be stable, and to compare favorably with the estimators introduced by Olkin et al. (1981) and Carroll and Lombard (1985).

DTIC COPY INSPECTED

Accesion For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Dist ibution /
Availability Codes
Dist | Avail and / or Special
A-1

Adrian E. Raftery is Associate Professor of Statistics and Sociology, University of Washington, Seattle, WA 98195. This work was supported by the Office of Naval Research under contract N00014-84-C-0169. I am grateful to W.S. Jewell for helpful discussions, and to Peter Guttorp for helpful comments on an earlier version of this paper.

1. INTRODUCTION

Suppose $x=(x_1,\ldots,x_n)$ is a set of success counts from a binomial distribution with unknown parameters N and θ. The problem of estimating N was first considered by Haldane (1942), who proposed the method of moments estimator, and Fisher (1942), who derived the maximum likelihood estimator. DeRiggi (1983) showed that the relevant likelihood function is unimodal. However, Olkin, Petkau, and Zidek (1981) - hereafter OPZ - showed that both these estimators can be unstable in the sense that a small change in the data can cause a large change in the estimate of N.

OPZ introduced modified estimators and showed that they are stable. On the basis of a simulation study, they recommended the estimator which they called MME:S. Casella (1986) suggested a more refined way of deciding whether or not to use a stabilised estimator. Kappenman (1983) introduced the "sample reuse" estimator; this performed similarly to MME:S in a simulation study, and is not further considered here. The history and applications of the problem were discussed in more detail by OPZ; a recent application was described by Dahiya (1980), who used the maximum likelihood estimator to estimate the population sizes of different types of organism in a plankton sample.

Draper and Guttman (1971) adopted a Bayesian approach, and gave a full solution for the case where N and θ are independent a priori, the prior distribution of N is uniform, and that of θ is beta. Blumenthal and Dahiya (1981) suggested N^* as an estimator of N, where (N^*,θ^*) is the joint posterior mode of (N,θ) with the Draper-Guttman prior. However, they did not say how the parameters of the beta prior for θ should be chosen. Carroll and Lombard (1985) - hereafter CL - recommended the N estimator Mbeta (1,1), the posterior mode of N with the Draper-

Guttman prior after integrating out θ, where the prior of θ has the form $p(\theta) \propto \theta(1-\theta)$ $(0 \leq \theta \leq 1)$.

Most of these papers were concerned almost exclusively with point estimation; interval estimation has been little studied. The simpler problem of estimating N when θ is known has been addressed by Feldman and Fox (1968), and Hunter and Griffiths (1978).

I adopt a Bayes empirical Bayes approach (Deely and Lindley 1981). This provides a simple way of specifying prior information, and also allows a convenient representation of vague prior knowledge using limiting, improper, prior forms. It leads to solutions of both the point estimation and interval estimation problems. The Bayes estimator corresponding to the relative squared error loss function and a vague prior distribution is shown in Section 3 to be stable, and, using simulation, to compare favorably with both MME:S and Mbeta (1,1).

2. A BAYES EMPIRICAL BAYES APPROACH

I assume that N has a Poisson distribution with mean μ. This defines an empirical Bayes model in the sense of Morris (1983). Then $x_1, \ldots, x_n$ are realisations of a Poisson random variable with mean $\lambda = \mu\theta$. I carry out a Bayesian analysis of this model.

I specify the prior distribution in terms of (λ,θ) rather than (μ,θ). This is because, if the prior is based on past experience, it would seem easier to formulate prior information about λ, the mean of the *observations*, than about μ, the mean of the *unobserved* quantity N. If this is so, prior information about λ would be more precise than that about μ or θ, so that it may be more reasonable to assume λ and θ independent a priori than μ and θ. In this case, μ and θ would be negatively associated a priori. Jewell (1985) has proposed a solution to the different but related problem of population size estimation from capture-recapture sampling, which is based on an

assumption similar to prior independence of μ and θ in the present context.

The posterior distribution of N is

$$p(N|x) \propto (N!)^{-1} \{\prod_{i=1}^{n} \binom{N}{x_i}\} \int_0^1 \int_0^\infty \theta^{-N+S} (1-\theta)^{nN-S} \lambda^N \exp(-\lambda/\theta) p(\lambda,\theta) \, d\lambda \, d\theta$$

$$(N \geq x_{max}) \qquad (2.1)$$

where $S = \sum_{i=1}^{n} x_i$, and $x_{max} = \max\{x_1, \ldots, x_n\}$. If λ and θ are independent a priori, and λ has a gamma prior distribution, so that $p(\lambda,\theta) \propto \lambda^{\kappa_1 - 1} e^{-\kappa_2 \lambda} p(\theta)$, then λ can be integrated out analytically, and (2.1) becomes

$$p(N|x) \propto (N!)^{-1} \Gamma(N+\kappa_1) \{\prod_{i=1}^{n} \binom{N}{x_i}\}$$

$$\int_0^1 \theta^{-N+S} (1-\theta)^{nN-S} (\theta^{-1}+\kappa_1)^{-(N+\kappa_2)} p(\theta) \, d\theta \qquad (N \geq x_{max})$$

I now consider the case where vague prior knowledge about the model parameters is represented by limiting, improper, prior forms. I use the prior $p(\lambda,\theta) \propto \lambda^{-1}$, which is the product of the standard vague prior for λ (Jaynes 1968) with a uniform prior for θ. This leads to the same solution as if a similar vague prior were used for (μ,θ), namely $p(\mu,\theta) \propto \mu^{-1}$. The posterior is

$$p(N|x) \propto \{(nN-S)!/(nN+1)!N\} \{\prod_{i=1}^{n} \binom{N}{x_i}\} \qquad (N \geq x_{max}) \qquad (2.2)$$

In the important special case where $n=1$, (2.2) becomes

$$p(N|x) = x_1/\{N(N+1)\} \qquad (N \geq x_1)$$

so that the posterior median is $2x_1$, which seems intuitively reasonable.

3. POINT ESTIMATION

Bayes estimators of N may be obtained by combining (2.2) with appropiate loss functions; examples are the posterior mode of N, MOD, and the posterior median of N, MED. Previous authors, including OPZ, CL, and Casella (1986) have agreed that the relative mean squared error of an estimator $\hat{N}$, equal to $E[(\hat{N}/N-1)^2]$, is an appropriate loss function for this problem. The Bayes estimator corresponding to this loss function is

$$\text{MRE} = \sum_{N=x_{max}}^{\infty} N^{-1}p(N|x) \Big/ \sum_{N=x_{max}}^{\infty} N^{-2}p(N|x)$$

The three Bayes estimators, MOD, MED, and MRE, are reasonably stable, as can be seen from the results for the eight particularly difficult cases listed in Table 2 of OPZ, which are shown in Table 1. MED was closer to the true value of N than the other estimators considered in four of the eight cases, while MOD was best in a further three cases. However, in the cases in which MOD was best, MED performed poorly; the converse was also true. The other three estimators always fell between MOD and MED.

Table 1 about here

The results of a simulation study are shown in Table 2. I used the same design as OPZ and CL. In each replication, N, θ, and n were generated from uniform distributions on [0,1],

$\{1, \ldots, 100\}$, and $\{3, \ldots, 22\}$ respectively, using the uniform random number generator of Marsaglia, Ananthanarayanan, and Paul (1973). A binomial success count was then generated using the IMSL routine GGBN. There were 2,000 replications.

Table 2 about here

Table 2 shows that MRE performed somewhat better than MME:S and Mbeta (1,1) in both stable and unstable cases, with an overall efficiency gain of about 10% over MME:S, and about 6% over Mbeta (1,1). Here, as in OPZ, a sample is defined to be stable if $\bar{x}/s^2 \geq 1+1/\sqrt{2}$, and unstable otherwise, where $\bar{x} = \sum x_i/n$, and $s^2 = \sum(x_i - \bar{x})^2/n$.

4. EXAMPLES

CL analyzed two examples, involving counts of impala herds and individual waterbuck. The point estimators are shown in Table 3. The stability of the Bayes estimators is again apparent; the stability of MRE for the waterbuck example is noteworthy given the highly unstable nature of this data set.

Table 3 about here

The posterior distributions obtained from (2.2) are shown in Figures 1 and 2. The posterior distribution for the waterbuck example has a very long tail; this may be related to the extreme instability of this data set.

Figures 1 and 2 about here

REFERENCES

Blumenthal, S., and Dahiya, R.C. (1981), "Estimating the Binomial Parameter n," *Journal of the American Statistical Association*, 76, 903-909.

Carroll, R.J., and Lombard, F. (1985), "A Note on N Estimators for the Binomial Distribution," *Journal of the American Statistical Association*, 80, 423-426.

Casella, G. (1986), "Stabilizing Binomial n Estimators," *Journal of the American Statistical Association*, 81, 172-175.

Dahiya, R.C. (1980), "Estimating the Population Sizes of Different Types of Organisms in a Plankton Sample," *Biometrics*, 36, 437-446.

DeRiggi, D.F. (1983), "Unimodality of Likelihood Functions for the Binomial Distribution," *Journal of the American Statistical Association*, 78, 181-183.

Deely, J.J., and Lindley, D.V. (1981), "Bayes Empirical Bayes," *Journal of the American Statistical Association*, 76, 833-841.

Draper, N., and Guttman, I. (1971), "Bayesian Estimation of the Binomial Parameter," *Technometrics*, 13, 667-673.

Feldman, D., and Fox, M. (1968), "Estimation of the Parameter n in the Binomial Distribution," *Journal of the American Statistical Association*, 63, 150-158.

Fisher, R.A. (1942), "The Negative Binomial Distribution," *Annals of Eugenics*, 11, 181-187.

Haldane, J.B.S. (1942), "The Fitting of Binomial Distributions," *Annals of Eugenics*, 11, 179-181.

Hunter, A.J., and Griffiths, H.J. (1978), "Bayesian Approach to Estimation of Insect Population Size," *Technometrics*, 20, 231-234.

Jaynes, E.T. (1968), "Prior Probabilities," *IEEE Transactions on Systems Science and Cybernetics*, SSC-4, 227-241.

Jewell, W.S. (1985), "Bayesian Estimation of Undetected Errors," *Theory of Reliability*, 94, 405-425.

Kappenman, R.F. (1983), "Parameter Estimation via Sample Reuse," *Journal of Statistical Computation and Simulation*, 16, 213-222.

Marsiglia, G., Ananthanarayanan, K., and Paul, N. (1973), *Super-Duper Random Number Generator*, Montreal: School of Computer Science, McGill University.

Morris, C.N. (1983), "Parametric Empirical Bayes Inference: Theory and Applications (with Discussion)," *Journal of the American Statistical Association*, 78, 47-65.

Olkin, I., Petkau, J., and Zidek, J.V. (1981), "A Comparison of n Estimators for the Binomial Distribution," *Journal of the American Statistical Association*, 76, 637-642.

Table 1. N Estimators for Selected and Perturbed Samples.

	Parameters			*Estimators*				
Sample	*N*	θ	*n*	MME:S	Mbeta (1,1)	MOD	MED	MRE
1	75	.32	5	70	49	42	82	57
				80	52	46	91	62
2	34	.57	4	77	47	42	84	57
				91	52	46	95	62
3	37	.17	20	25	23	21	40	26
				27	25	23	46	29
4	48	.06	15	10	8	7	14	10
				12	10	10	19	12
5	40	.17	12	26	25	23	42	30
				32	29	27	52	35
6	74	.68	12	153	125	114	207	127
				162	131	120	217	129
7	55	.48	20	69	63	59	91	75
				74	67	63	101	81
8	60	.24	15	49	41	38	68	49
				53	45	41	77	53

NOTE: The exact samples are given in Table 2 of OPZ. For each sample number, the first entries are the *N* estimates for the original sample, and the second entries are the *N* estimates for the perturbed sample obtained by adding one to the largest success count.

Table 2. Relative Mean Square Errors of the N Estimators

Cases	*No.*	*Estimators*		
		MME:S	Mbeta (1,1)	MRE
All cases	2000	.171	.165	.156
Stable cases	1378	.108	.104	.100
Unstable cases	622	.312	.300	.281

Table 3. Estimators for the Impala and Waterbuck Examples: Original and Perturbed Samples

	Estimators				
Example	MME:S	Mbeta (1,1)	MOD	MED	MRE
Impala	54	42	37	67	49
	63	46	40	76	54
Waterbuck	199	140	122	223	131
	215	146	127	232	132

NOTE: The data are given in Section 4 of CL. For each example, the first entries are the N estimates for the original sample, and the second entries are the N estimates for the perturbed sample obtained by adding one to the largest success count.

Figure 1

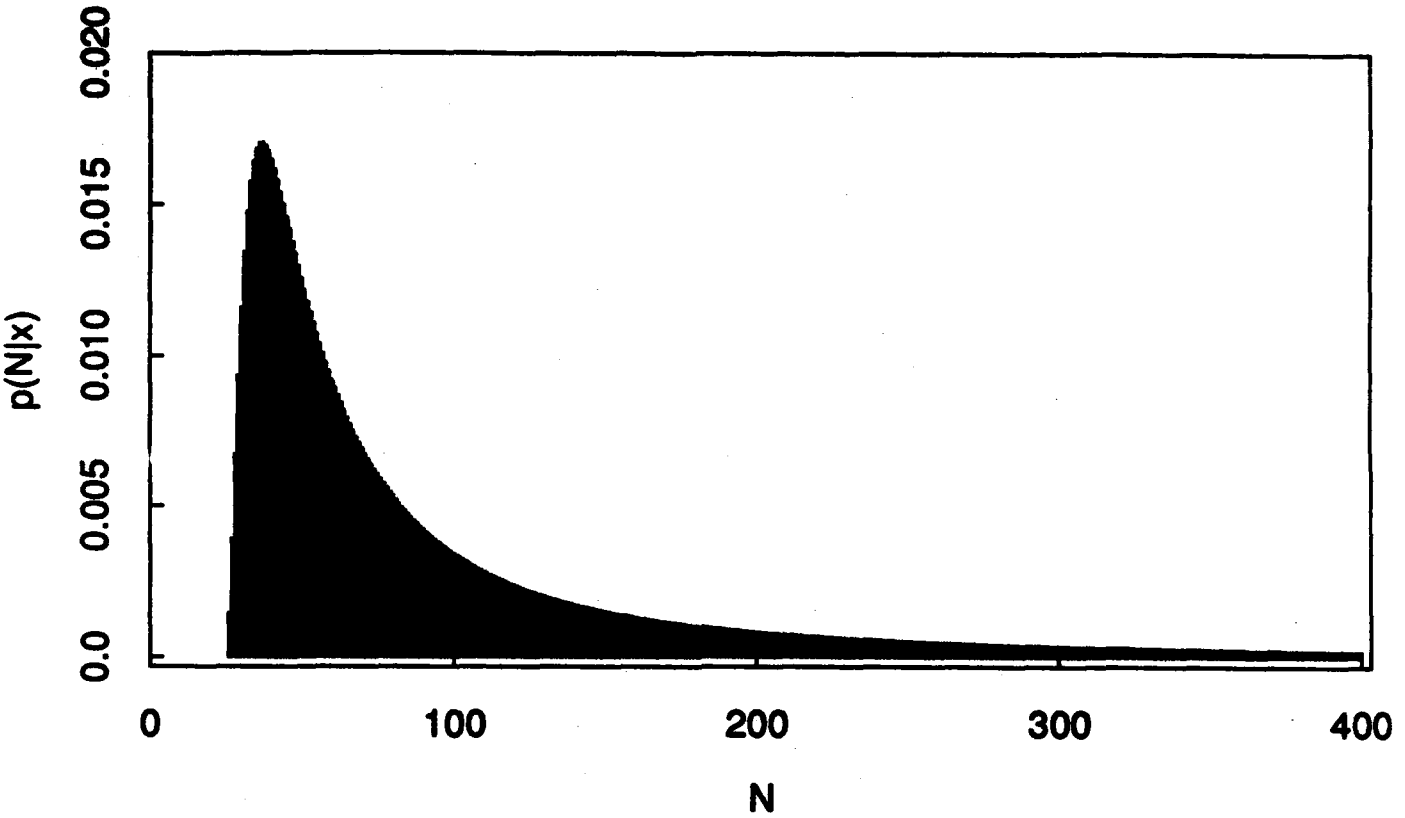
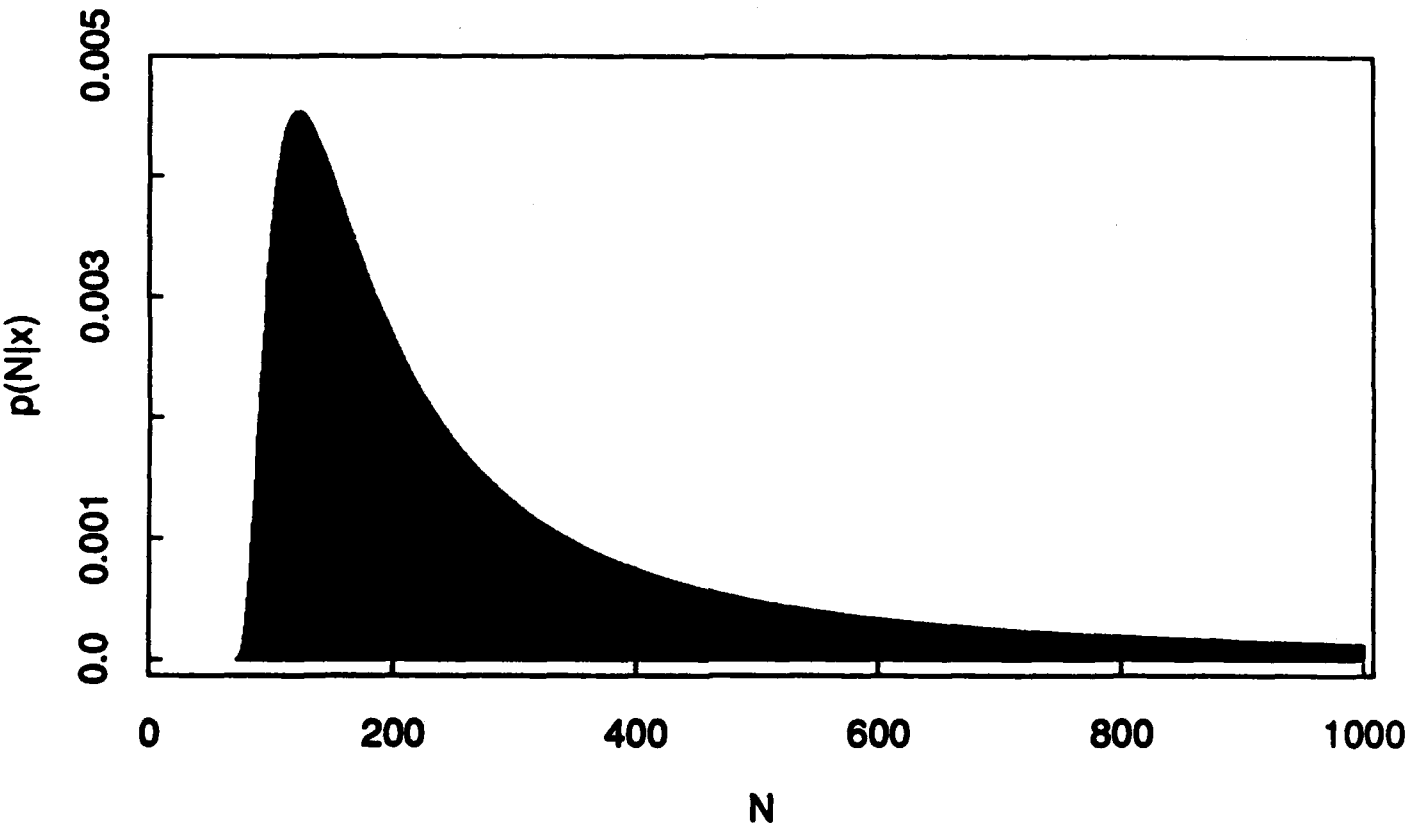


Figure 2



END

7-87

DTIC